263-2013

Do people still miss Steve Jobs as the CEO of Apple Inc.? A Text Mining Approach: Comparing SAS® and R

Anurag Srivastava and Pranav Karnavat, Shanti Business School, Ahmedabad, India

Guide: Prof. Amit Saraswat

ABSTRACT

Social media is considered as a powerful tool to provide valuable information about views, expressions, need and expectation of people. Marketers need such information to capitalize upon and satisfy needs and expectation of the consumers. Twitter is a powerful social media website which had over 500 million users worldwide as on 9th may 2012 (Twitter Inc., 2012). Tweets posted can be analyzed to get insight about relationships and patterns hidden inside the textual data. In this paper tweets are collected about Steve Jobs – post his sad demise to find out what customers think of Apple Inc., now using text mining technique in SAS and R. 'Get tweet' macro is used to fetch data from twitter in SAS while 'twitteR' package is used to fetch data from twitter in R. SAS Text Miner was used in SAS to analyze the data while 'tm' package was used to analyze the data in R.

INTRODUCTION

In today's fast paced world the environment has become all the more dynamic due to consumers being affected by an assortment of internal and external factors of the market and hence it becomes very important for a marketer to capture and analyze real time information contained in data. It is also often observed that a respondent is restricted to respond on a fixed number of dimensions through a questionnaire which may lead to the full information from the respondent – not being revealed which also is not real time information (Bloomberg, 2012). Vast amounts of new information and data are generated everyday through economic, academic and social activities. This sea of data, predicted to increase at a rate of 40% p.a., has significant potential economic and societal value. Techniques such as text are required to exploit this potential. Social networking sites such as Facebook and Twitter enable users to share over 1.3 billion pieces of information/content per day (McDonald and Kelly, 2012). Social Media is now considered as a valuable source to monitor customers or public views, emotions and expectations as customers freely express their views on various topics and issues and this real time information can be fetched from social media sites through text mining. Thus, if the textual data is analyzed, social media can prove to be a rich source of potential information. Information available in text is clear and explicitly stated on various dimensions at respondent's will. The information might produce relationships or patterns that are buried in the text and which would otherwise be difficult, if not impossible, to discover with data mining as those dimensions may or may not be included. Text mining technique is the solution for information retrieval, processing and analyzing text information.

Twitter is considered as a powerful social media tool to detect patterns, trends, moods and sentiments of users over a topic or an issue (Garla and Chakraborty, 2012). It holds more than 500 million users around the globe (Twitter Inc., 2012).

Steven Paul "Steve" Jobs (February 24, 1955 – October 5, 2011) was an American entrepreneur. He is best known as the co-founder, chairman, and CEO of Apple Inc. Through Apple, he was widely recognized as a charismatic pioneer of the personal computer revolution. Steve Jobs was named as one of the 20 most influential Americans in July 2012 (Time Magazine, 2012). His name was placed among the iconic greats like George Washington, Alexander, Graham bell and Albert Einstein. Credit to the goodwill which Apple Inc. has earned goes to Steve jobs (Isaacson, 2011). Consumers feel proud in owning an Apple product. Not only purchasing Apple products but consumers are also loyal to Apple Inc. It would not be unfair to say that Steve Jobs played a major role in Apple's success, because following this return in 1997, Apple Inc. again witnessed a rise in credibility and as a brand like never before, which earlier had neared to bankruptcy when he was removed from Apple Inc. from 1985 – 1997 (Bloomberg Business Week Magazine, 2011). His sad demise was a day of mourning for the followers and loyal customers of Apple Inc. People all over the world expressed their condolences and sorrows by posting about him on blogs and social networking sites. Decease of Steve Jobs created a lot of buzz on Twitter Inc. with so many people tweeting about him.

This paper tries to find out if people still miss Steve Jobs as the CEO of Apple Inc.

OBJECTIVES

The objective of the paper is –

1. To apply text mining technique and analyze tweets regarding Steve Jobs.
2. To understand whether people still miss Steve Jobs as the CEO of Apple Inc.
3. To identify the top of mind recall terms related to Steve Jobs using Concept Link Diagram.
4. To compare the results of text mining technique executed on SAS® and R.

LITERATURE REVIEW

Text mining is a comprehensive technique describing a range of technologies for analyzing and processing semi structured and unstructured textual data. The integrated premise behind each of these technologies is the need to convert text into numbers so that powerful analytical algorithms can be applied to large document databases. This requires knowing how to both use and combine techniques for handling text, ranging from individual words to documents to corpus which is the entire document database. The purpose of text mining is to derive high level information which is a combination of relevance, novelty and insight through statistical pattern learning (Weiss et al., 2005). Specific grammar rules and language conventions govern the way a language is used leading to statistical patterns appearing frequently in large amount of text. It is true that maximum amount of business relevant information originates in an unstructured form, primarily text and hence syntax is used in text mining which pertains to a set of grammatical rules and the structure of language used. Further, the circumstance and the usage of the terms has to be determined which is done by semantics which refers to the meaning of individual words within surrounding context (Miner et al, 2012). Syntax and semantics go hand in hand to complete the meaning of a sentence. There are seven different areas of applications of text mining – Search and Information Retrieval, Document Classification, Information Extraction, Natural Language Processing, Concept Extraction, Document Clustering and Web Mining. These areas include processes like searching textual data, retrieving and extracting information, classifying and clustering the documents into similar sets and extracting concepts from the textual data available. Also data available on internet can also be analyzed which is known as web mining.

BENEFITS OF TEXT MINING

Text mining and analytics have the potential to increase the research base available to business and society and to enable business and others to use the research base more effectively. Few of the benefits of text mining are as follows

RECOGNIZES TRENDS AND BUSINESS OPPORTUNITIES

Text Mining can transform unstructured text into numeric representations that surmise the collection. This data then can become insightful input to full range of predictive and data mining modeling techniques. In turn, marketer can better understand customer, service and product needs – and predict opportunities for timely exploitation.

INCREASED RESEARCH EFFICIENCY

A key benefit of text mining is that it enables much more efficient analysis of extant knowledge. The ability to extract information automatically cuts down the time spent on ensuring coverage of domain knowledge in the literature review process. For example, given the sheer volume of scholarly publications now available, it could take a human researcher several years to analyze and identify all relevant sources for a particular problem. Using text mining, identifying relevant material could drastically cut down the time required.

UNLOCKING HIDDEN INFORMATION

As there are enormous amount of academic publications and literature then there is probability that there may be underlying connections between different subtopics which can be understood only with automated analysis.

BROADER ECONOMIC AND SOCIAL BENEFITS

- Text mining helps in reducing cost incurred in doing exploratory research and can help in gaining productive gains.
- There is potential for new radical and incremental innovation with wider economic benefit and including innovative service development (McDonald and Kelly, 2012).

STEPS IN TEXT MINING

PRE-PROCESSING TEXT

Preprocessing is converting the text to structured format to make it ready for analysis which contains the below stated steps.

1. **Tokenization**:
   Tokenization is the process of giving numerical identity (tokens) to the textual words. Thus the textual data is broken into distinct tokens.

2. **Removing Stop Words and stemming**
   Stop words are the words which carry no meaning but are actually in the sentence so that the sentence is grammatical correct. Analyzing them would not add up any value. Hence such words are not considered in the analysis. Stemming, in analytical terms, is pruning. It removes prefixes and suffixes attached to the root word to get the real term in the data.
3. **Normalize Case**
   This includes converting the text to either lower or upper case for uniformity in the data so that further analysis becomes easy. (Miner et al, 2012,)

PROCESSING TEXT

Creating vectors from text

Vectors need to be created as it is necessary for text mining algorithms to work. Vector representation can take one of the three different forms - (1) Binary Representation, (2) Integer Count or (3) Float-Valued Weighted Vector. (Albright, 2004)

Creating Term by Document Matrix

Term by Document matrix is a matrix which explains number of terms present in each document. This matrix becomes base for weighted term by document matrix.

Singular value decomposition (Latent Semantic Analysis)

Singular value decomposition reduces number of dimensions of the data. Similar dimensions are clubbed together so that data becomes easier to read and analyze (Albright, 2004).

HYPOTHESIS

Based on the above understanding, this paper proposes the below mentioned hypothesis and its alternate.

H0: People still miss Steve Jobs as the CEO of Apple Inc.

H1: People do not miss Steve Jobs as the CEO of Apple Inc.

RESEARCH METHODOLOGY

In this paper, text mining has been done using two tools – SAS® and R.

TEXT MINING IN SAS®

In SAS® the macro 'Get tweet' collects customize data from Twitter. All the Tweets that match the search conditions were downloaded as a SAS data set. The keyword parameters specified in the macro were 'Steve Jobs'. The macro does a basic data cleaning such as removing "http", tags, URLs, and stripping of characters in ID variables. Text Miner node available in SAS® Enterprise Miner is used for analysis. It gives weights to terms based on their frequency. If a particular term appears more often, it is given a lower weight. Certain parts of speech were ignored for appropriate parsing – conversion of textual data, ready for analysis, to tokens. Also punctuation marks were ignored during the analysis as they do not add meaning to the analysis. Parsing for numbers & entities – organization's names, people's names, product names, locations, dates, measurement, currency and time, was allowed so that the entire document is classified into different entities and appropriate weights are assigned to terms and numbers based on their frequency of occurrence. Enterprise miner can classify terms among all the parts of speech.. Here conjunctions, determiners, pronouns, abbreviation and participles were ignored as they were low content terms.

TEXT MINING IN R

Text Mining in R consisted of the following steps:

1. Download and install twitteR package to fetch data from twitter.
2. Download and install tm package to apply text mining technique on the data collected
3. Fetch data from twitter which can be done using following code:

```
R <- searchTwitter('Steve Jobs', since= '2012-11-03', until= '2012-11-17', n=1500)
R<-Corpus(DirSource("E:/r data"))
R <- tm_map(R, removeNumbers)
R <- tm_map(R, removePunctuation)
R <- tm_map(R, tolower)
R <- tm_map(R, stemDocument, language = "english")
R <- TermDocumentMatrix(R)
findFreqTerms(mydata.dtm, lowfreq=30)
```

OUTPUT AND ANALYSIS

OUTPUT IN SAS®

Terms

Term	Role	Attribute	Freq	# Documents	Keep	Weight
steve	Prop	Alpha	553	547	Y	0.01459
steve jobs	PERSON	Entity	497	495	Y	0.02981
jobs	Prop	Alpha	487	485	Y	0.033
rt	Prop	Alpha	168	167	Y	0.19987
+ apple	Noun	Alpha	163	161	Y	0.20596
+ make	Verb	Alpha	400	150	Y	0.22255
gates	Prop	Alpha	148	148	Y	0.2184
bill	Prop	Alpha	147	147	Y	0.21946
bill gates	PERSON	Entity	144	144	Y	0.22269
microsoft	COMPANY	Entity	123	123	Y	0.24734
+ not	Adv	Alpha	114	112	Y	0.26303
+ do	Verb	Alpha	105	101	Y	0.28035
+ job	Noun	Alpha	89	88	Y	0.30038
mallette	Prop	Alpha	83	83	Y	0.30886
pattie	Prop	Alpha	83	83	Y	0.30886
pattie malle	PERSON	Entity	83	83	Y	0.30886
in	Prep	Alpha	83	80	Y	0.3167
justin	TITLE	Entity	69	69	Y	0.33776
time	Noun	Alpha	64	64	Y	0.34952
+ will	Aux	Alpha	64	58	Y	0.36985
dabieberbu	Noun	Alpha	57	57	Y	0.36764
life	Noun	Alpha	57	57	Y	0.36764
+ have	Verb	Alpha	74	57	Y	0.37663
+ live	Verb	Alpha	55	54	Y	0.37717
+ limit	Verb	Alpha	52	52	Y	0.382
jay	Noun	Alpha	38	38	Y	0.43106
jay tomlins	PERSON	Entity	38	38	Y	0.43106
louis	Noun	Alpha	38	38	Y	0.43106
louis tomlin	PERSON	Entity	38	38	Y	0.43106
peasant	Noun	Alpha	38	38	Y	0.43106
sorkin	Prop	Alpha	38	38	Y	0.43106
tomlinson	Prop	Alpha	76	38	Y	0.43106
+ look	Verb	Alpha	40	35	Y	0.45014
aaron	Prop	Alpha	33	33	Y	0.45312
justelounor	Prop	Alpha	33	33	Y	0.45312
+ quote	Noun	Alpha	33	33	Y	0.45312
+ job	Verb	Alpha	34	32	Y	0.46121
else	Adj	Alpha	28	28	Y	0.47882
now	Adv	Alpha	28	28	Y	0.47882

Figure 4: Partial Output of Terms matrix

Figure 4 summarizes the information in the data. It displays terms in the data, its parts of speech, frequency and weight of the word. It is seen in the output that term 'Steve' has received lowest weight as it has highest frequency.

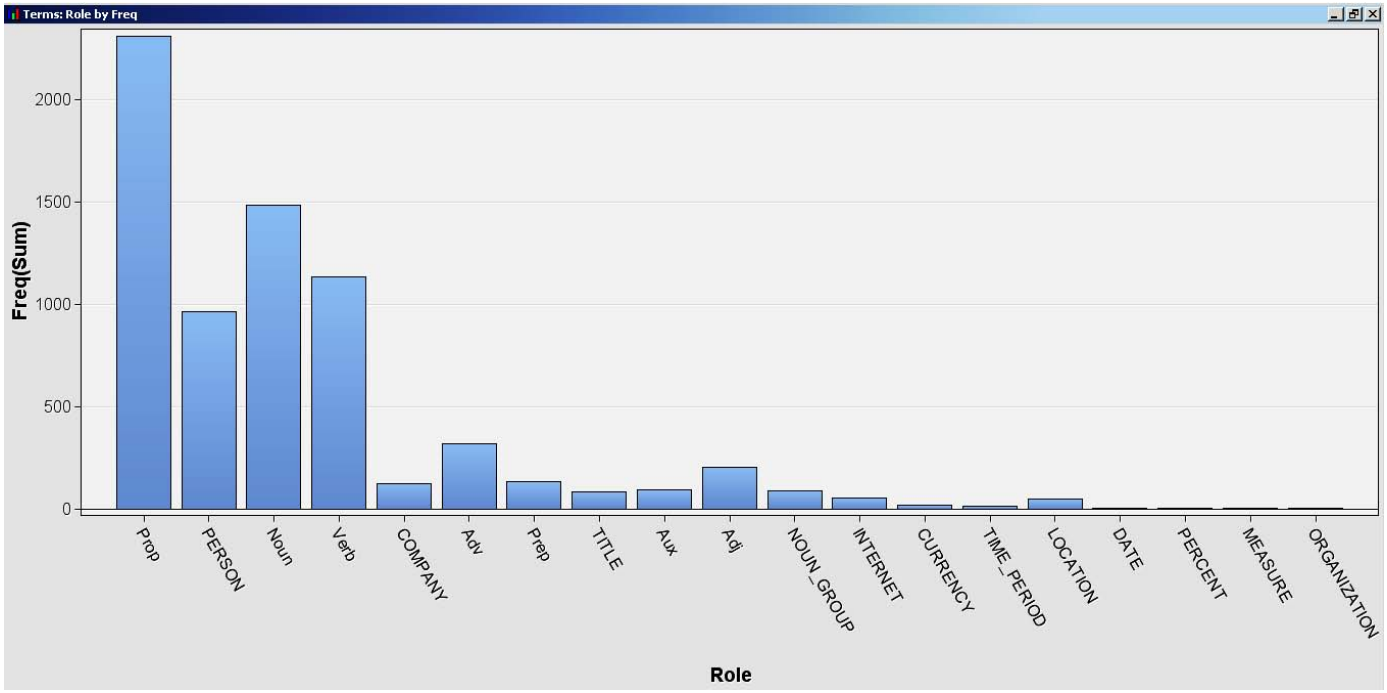

Figure 5: Output of Role by frequency matrix

Figure 5 shows frequency of the terms according to the parts of speech and entity of the terms. For example proper noun has highest frequency while the organization has very less frequency.

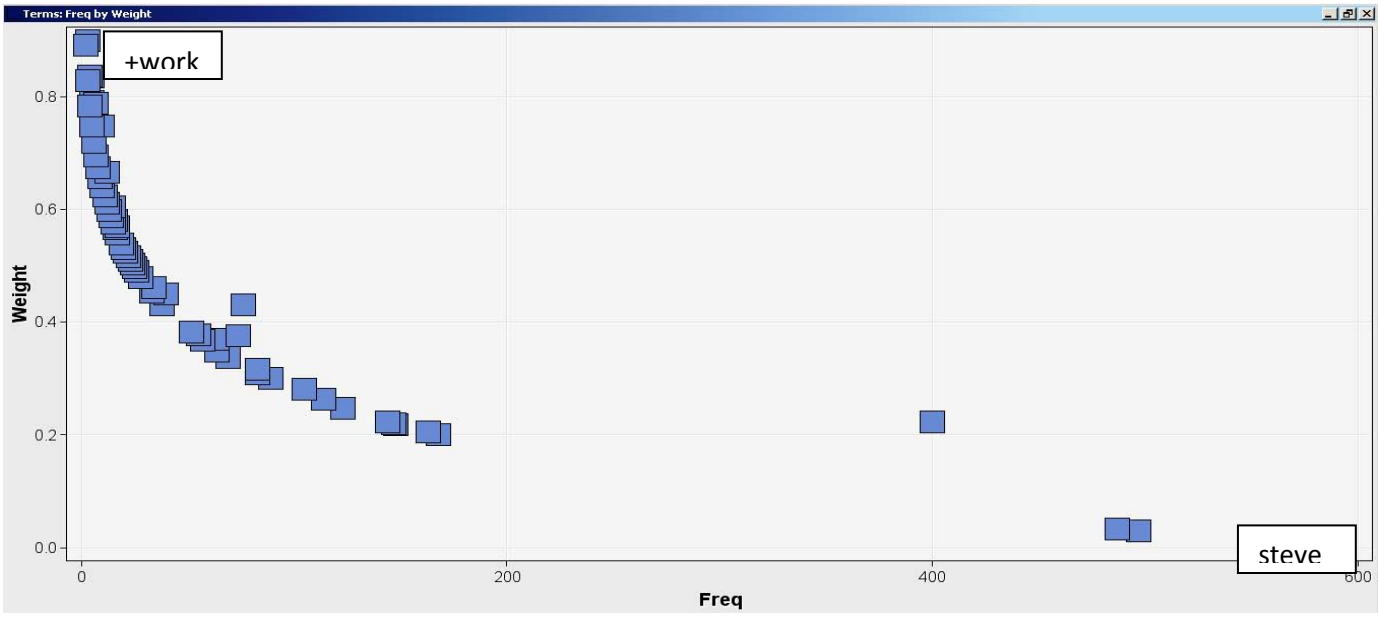

Figure 6: Output of Frequency by Weight

Figure 6 shows a plot diagram of the terms keeping frequency and weight on X and Y dimension respectively. It shows that the word “Steve” is repeated many times and has got less weightage while work has very less frequency and high weightage. Each blue box is a term which is anlayzed.

ANALYSIS IN SAS®

CONCEPT LINK DIAGRAM

Concept linking of terms shows strength/weakness of relationship between all the terms mentioned in the dataset. Interconnections between the terms are well understood through concept links. Strength of the term with other term is determined by width of the link which joins those terms. If the width of the link to the term is more the relationship between the between the terms is strong. The interpretation here cannot be done on the basis of seeing some value which can be done in case of data mining techniques. Interpretation requires in depth understanding of the topic, logical sense and updated knowledge about market scenario. Since the software analyzes on the basis of frequencies and weights it is not certain that terms may also be semantically close to each other. Also few terms revealing strong relationship in the concept link diagram may have weak relationship in the real world. However there are rare possibilities of vice versa being true. Hence the interpretation can be made after gaining sufficient knowledge of the domain as well as text mining technoque. In the paper, Concept links are performed on words such as 'Steve jobs' and 'Death'.

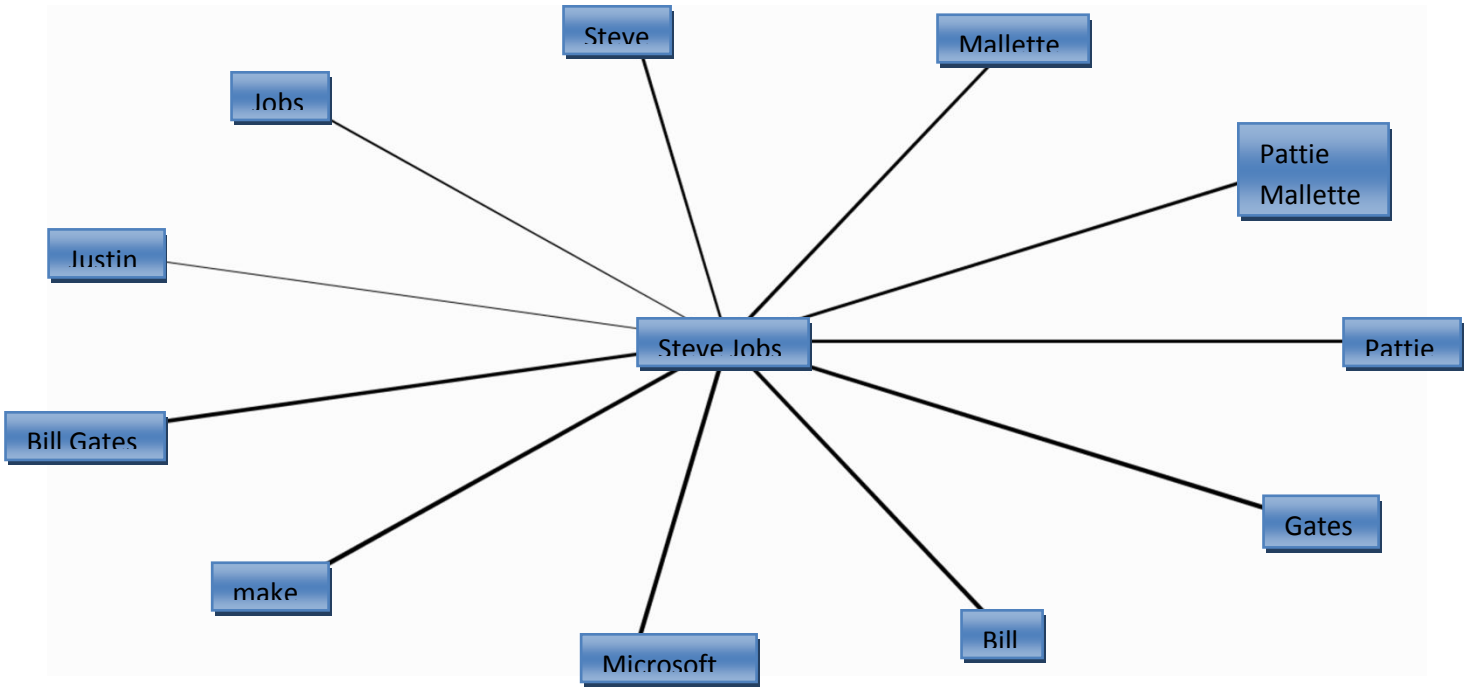

Figure 7: Concept linking on 'Steve Jobs'

Figure 7 depicts that the term "Steve Jobs" is linked strongly with Bill Gates, Microsoft, Make, Mallete and few terms. The result is not explicitly stated for a layman to interpret it. A mix of Logical and business and environment knowledge is pre-requisite for the output to understand. Figure 7 output can be understood in way that is stated as under

1. Bill gates and Steve jobs are makers of (apple) and Microsoft.
2. There was a joke which was tweeted and retweeted many times in which Steve Jobs has said that he is creator of Apple Inc., Bill Gates has said that he is creator of Microsoft and Pattie Mallete (famous writer) said that she is mother of Justin Beiber (singer and musician).
3. Other terms should be ignored because those are the first names and last names of individuals already analyzed.

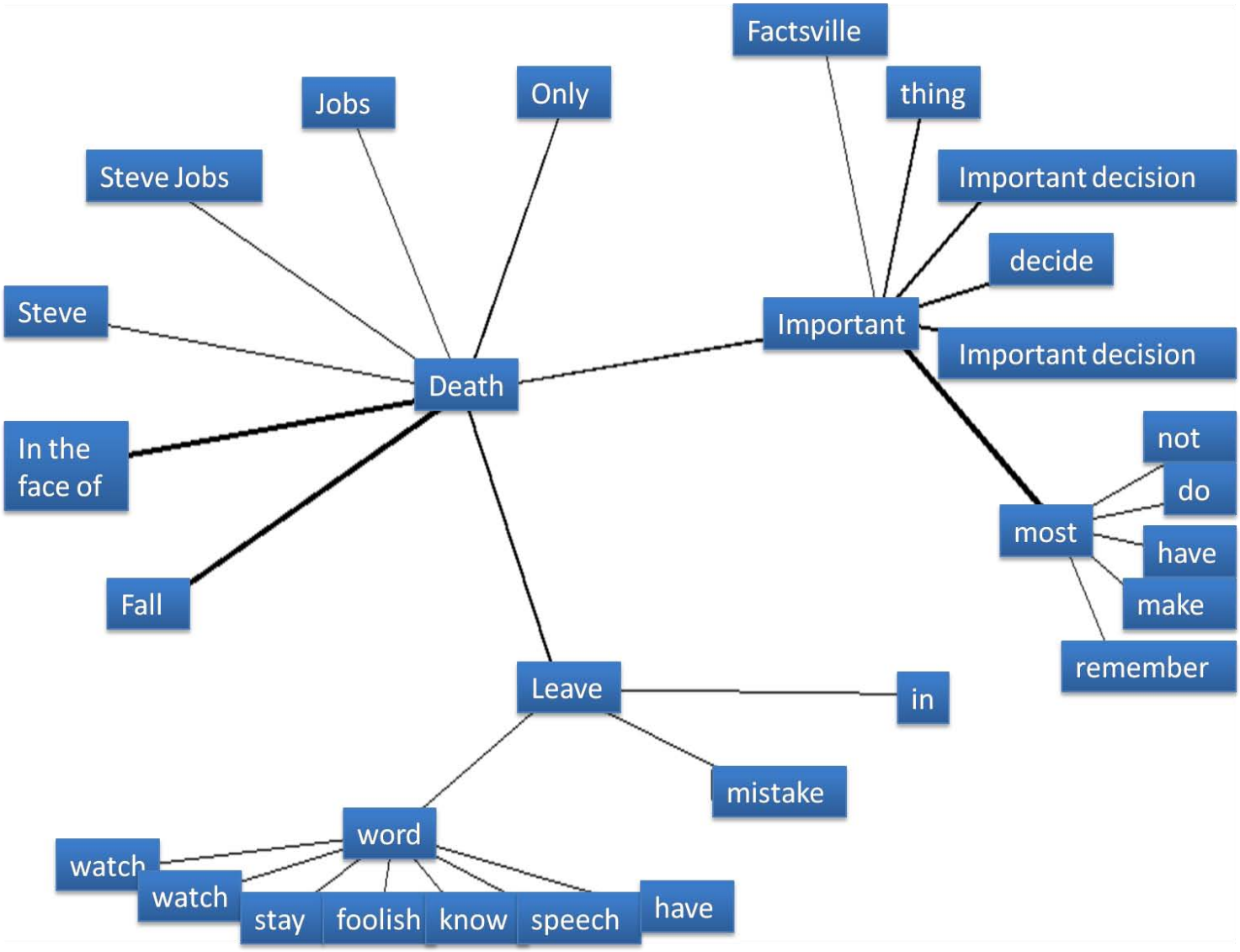

Figure 8: Concept linking on 'Death'

Terms like Steve Jobs, important, decision, speech, remember, leave are the words which are closely related to death of Steve Jobs. These words are talking about many dimensions which could only have been understood through text mining. Concept linking reveals that the public in general is of the belief that –

- Steve Jobs took a very important decision leaving the company (before death).
- Few also say that it was a mistake.
- His last words include stay hungry, stay foolish.
- He would always be remembered.
- In this case also, unnecessary links should be avoided as they do not create any meaning

OUTPUT IN R

TERM DOCUMENT MATRIX

A term document matrix is created after analyzing the tweets fetched from twitter as shown in Figure 9. It summarizes the data by creating a term by document matrix, mentioning number of terms in the data in each document which were parsed.

```
A term-document matrix (4467 terms, 1 documents)

Non-/sparse entries: 4467/0
Sparsity           : 0%
Maximal term length: 140
Weighting          : term frequency (tf)
```

Figure 9 Term Document Matrix output in R

LIST OF FREQUENT TERMS

The code 'Find freq terms' reveals Frequent terms which appear maximum times in the tweets related to Steve Jobs. It is observed that words like apple, steve, laptop, iphone, bill gates appear frequently and thus it is understood that they are used maximum times together. The output is similar to the output of SAS wherein same terms had highest frequency as shown in Figure 10.

```
 [1] "<U+393C><U+3E33>be"        "<U+393C><U+3E37>steve"     "aaron"
 [7] "and"                       "appl"                      "arent"
[13] "biopic"                    "bob"                       "cash"
[19] "day"                       "del"                       "desk"
[25] "excel"                     "expected<U+393C><U+3E34>"  "f<U+663C><U+3E61>tbol"
[31] "gate"                      "gioniza"                   "groovboard"
[37] "his"                       "hope"                      "introduc"
[43] "johnni"                    "just"                      "kevin"
[49] "life"                      "like"                      "live"
[55] "movi"                      "nnnwhen"                   "now"
[61] "pleas"                     "qualiti"                   "que"
[67] "sobr"                      "some"                      "sorkin"
[73] "that"                      "the"                       "this"
[79] "via"                       "was"                       "what"
[85] "yardstick"                 "year"                      "you"
"stay"                           "steve"                     "terminal<U+663C><U+3E31>"
"three"                          "use"                       "valdano"
"where"                          "will"                      "world"
"your"
```

Figure 10

COMPARING SAS® AND R

Both the software fetches 1500 tweets and fetches same dataset. However, R is limited to displaying only the frequencies of terms whereas SAS® creates Concept Link Diagram showing relationships and strength between terms, which is a step further in the insight obtained. SAS® classifies the terms in different entities whereas there is no scope for classification of terms in R. Apart from classification SAS® also classifies different terms according to their similarity with the document sets.

The final version of the paper will contain analysis of tweets of past two years - 1 year prior to and 1 year post Steve Jobs' demise using a SAS® Macro that will be created to fetch two years' data. This will be done using a campaign that will be created with the help of an expert where public can express and pool their feelings about the specified topic in the form of unstructured textual data. This data will then be finally analyzed and a comparative study of SAS and R will be reported based on the output, of text mining technique.

REFERENCES

Miner et al, G, 2012, Practical Text Mining and Statistical Analysis for Non Structured Text data application, 1st edition, USA, Academic Press.

Weiss et al., S, 2005, Predictive methods for analyzing unstructured information, USA, Springer.

Taming Text with the SVD, Russ Albright, SAS Institute Inc, 2004

%GetTweet: A New SAS® Macro to Fetch and Summarize Tweets, Satish Garla and Goutam Chakraborty, Oklahoma State University, SAS, 2012.

Jodi Blomberg, Twitter and Facebook Analysis: It's Not Just for Marketing Anymore, SAS, 2012.

Kathy Lange and Saratendu Sethi, What are people saying about your company, your products, or your brand?, SAS Institute Inc.

The Value and Benefit of Text Mining, Diane Mcdonald and Ursula Kelly, JISC, 2012.

Package 'twitteR', Jeff Gentry, February 20, 2012.

Package 'tm', Ingo Feinerer, February 15, 2012.

ACKNOWLEDGEMENT

It was a stimulating and motivating experience in completing this paper. It took lot of endeavor, time and energy. Several people we would like to thank for supporting us through this process. We would specially like to thank our guide, Prof. Amit Saraswat, Faculty – Decision Sciences at Shanti Business School, Ahmedabad, Gujarat, India. The supervision and support that he gave us truly helped in delivering quality content in the paper. The co-operation is indeed appreciated.

We appreciate and acknowledge contribution of Miss. Sadhana Singh and Mr. Vikram Suklani in helping us with programming for Text Mining in SAS.

CONTACT INFORMATION

Pranav Karnavat

pravkarnavat@gmail.com

Anurag Srivastava

anurag_srivastava@de-quo.com

ANNEXURE 1

SAS code for fetching data from Twitter and converting it to SAS dataset.

```
%macro gettweet(WORDS=,PHRASE=,ANY=,NONE=,HASH=,FROM=,TO=,SINCE=,
    UNTIL=,QUESTION=,CODE=,PATH=);
libname Twit "&PATH";
%let dataset=Tweets; /*Give a name for Destination Data set*/
filename httpOut "&PATH\twitterOutput.xml";
filename hOut "&PATH\httpOutputHeaders.txt";
filename hIn temp;
filename httpreq temp;
/*Add "+" between keyword parameters when Multiple words are used in the search*/
%let WORDS=%sysfunc(translate(%sysfunc(strip(&WORDS)),"+"," "));
/*%let PHRASE=%sysfunc(translate(%sysfunc(strip(&PHRASE)),"+"," "));*/
/*%let ANY=%sysfunc(translate(%sysfunc(strip(&ANY)),"+"," "));*/
/*%let NONE=%sysfunc(translate(%sysfunc(strip(&NONE)),"+"," "));*/
/*%let HASH=%sysfunc(translate(%sysfunc(strip(&HASH)),"+"," "));*/
/*%let FROM=%sysfunc(translate(%sysfunc(strip(&FROM)),"+"," "));*/
/*%let TO=%sysfunc(translate(%sysfunc(strip(&TO)),"+"," "));*/
%if &QUESTION=1
%then
%let QUESTION=%nrstr(&tude[]=%3F);
%else
%let QUESTION=%nrstr(&tude[]=);
%let search=%nrstr(q=&ands=)&WORDS%nrstr(&phrase=)&PHRASE%nrstr(&ors=)
&ANY%nrstr(&nots=)&NONE%nrstr(&tag=)&HASH%nrstr(&lang=en)%nrstr(&from=)&FROM%nrstr(&to
=)&TO%nrstr(&since=2012-11-03)&SINCE%nrstr(&until=2012-11-17)&UNTIL&QUESTION;
/*Create a Temp file used in PROC HTTP headerin option to hold base64 encode*/
data _null_;
file hIn;
put &code;
```

```
run;
/*Create Destination Data set*/
proc sql;
create table TWIT.&dataset
(
  id char(39) format=$39. informat=$39.,
  published num format=IS8601DT19. informat=IS8601DT19.,
  title char(159) format=$159. informat=$159.,
  updated num format=IS8601DT19. informat=IS8601DT19.,
  twitter_source char(100) format=$100. informat=$100.,
  twitter_lang char(2) format=$2. informat=$2.,
  uri char(50) format=$50. informat=$50.,
  content char(2600) format=$2600. informat=$2600.
); quit;
/*Initialize and increment Page Number*/
%let pageno=1;
%StartLoop:
/*Define a variable for the number of tweets per page. The maximum allowed is 100 */
%let pagerate=%nrstr(&rpp=100&page=);
/*Combine Search String, Page rate and Page Number Macro variables*/
%let searchstring="&search&pagerate&pageno";
/*Create a Temp file used in PROC HTTP IN= option */
data _null_;
file httpreq;
put &searchstring;
run;
proc http
     in=httpReq
out=httpOut
headerin=hIn
```

```
headerout=hOut
url="http://search.twitter.com/search.atom"
     method="get"
     ct="application/x-www-form-urlencoded";
run;
/*Define XML Mapper and XML Library*/
filename SXMLMAP "&PATH\TwitterSearch.map";
filename XMLLib "&PATH\twitterOutput.xml";
libname XMLLib xml xmlmap=SXMLMAP ACCESS=readonly;
/*Concatenate the XML Results in „Entry? data set and destination data sets*/
data twit.&dataset;
set twit.&dataset XMLLib.entry;
run;
/*Query the count of Tweets returned, into the Macro Variable „obscount?*/
proc sql noprint;
select count(*) into :obscount from XMLLib.entry;
quit;
%let pageno=%eval(&pageno+1); /*Increment Page Number*/
/**The Loop terminates if it is Page Fifteen or If it is the Last Page (<15) and has
less than 100 tweets. We can fetch a maximum of 1500 tweets at a time. If the tweets
available are less than 1500 the loop is terminated else the tweets from the last
fetched page keep writing to the Data set**/
%if %eval(&pageno)=16 or %eval(&obscount)<100
%then %goto EndLoop;
%else %goto StartLoop;
%EndLoop:
/*Summarize Tweets*/
proc sort data=twit.&dataset out=&dataset._temp nodupkey;
by title uri; run;    /*Delete duplicates*/
/*Below DATA step cleans tweets and creates two data sets (one with all the tweets and
```

```
the other only with retweets) in the work library*/

data &dataset (KEEP= id pubdate text author source retweet)

&dataset._rt (KEEP= id pubdate title text author source tweet_owner);

set &dataset._temp;

length text $ 159 tweet_owner $ 20 source $ 20;

format pubdate date7.;

retweet=0;

text=title;

author=tranwrd(uri,'http://twitter.com/','@');

id=substr(id,29);

pubdate=datepart(published);

if substr(text,1,3)='RT ' then do;

retweet=1;

tweet_owner=tranwrd(scan(text,2),':','');

call scan(text, 3, position, length);

text=substr(text,position);

end;

if _n_=1 then do;

retain pattern pattern2;

pattern = PRXPARSE ("s/(RT @.[^ ]*)|((http|www)(\d|\D)[^ ]*)|(@.[^ ]*)//i");

pattern2 = PRXPARSE ('/"(\w|\W)[^ ]*"/');

end;

call prxchange(pattern,-1,text);

text=strip(text);

if prxmatch(pattern2, twitter_source) then do;

call prxposn(pattern2, 0, position, length);

       source = strip(substr(twitter_source, position+8, length-9));

end;

if retweet=1 then output &dataset._rt;

output &dataset;
```

```
run;
/*Calcualte Total Number of records collected from twitter */
proc sql noprint;
select count(*) into :twtcount from twit.&dataset;
quit;
proc sql noprint;
select count(*) into :retwtcount from &dataset._rt;
quit;
/*Terminate macro execution if usernames are specified in FROM= parameter.
No Tweet report is generated. Only Data sets are created */
%if %length(&FROM) ne 0 %then %goto EndMacro;
proc sql outobs=10;
create table toptweeters as
select author 'Tweeter',count(author) 'Tweets'
from &dataset
group by author
order by 2 desc;
quit;
proc sql outobs=10;
create table topsources as
select source 'Source',count(source) 'Count'
from &dataset._rt
group by source
order by 2 desc;
quit;
proc sql outobs=10;
create table topowners as
select tweet_owner 'Tweeter',count(tweet_owner) 'Count'
from &dataset._rt
group by tweet_owner
```

```
order by 2 desc;
quit;
/*Define ODS Layout and generate PDF Report*/
options orientation=landscape;
goptions reset=all dev=sasprtc ftext="Helvetica";
ods listing close;
ods pdf file="&PATH\TweetReport_&dataset..pdf" STARTPAGE=NO BOOKMARKGEN=NO;
ods layout start;
ods region x=0 in y=0 in height=8.5 in width=11 in;
proc gslide;
title1 h=17pt j=Center underlin=1 'Tweet Report'  lspace=.1in;
title2 h=12pt j=Left " Tweets:&twtcount" " Retweets:&retwtcount"
lspace=.1in;
run;quit;
goptions border;
ods region x=0.25 in y=0.3 in height=3.5 in width=5 in;
title1;
axis1 label=(angle=90 "Tweeter")  minor=none;
axis2 label=(height=15pt "Top 10 Tweeters") minor=none;
proc gchart data=toptweeters;
hbar author/ sumvar= _TEMA001 maxis=axis1 raxis=axis2;
run;quit;
axis1 label=(angle=90 "Source")  minor=none;
axis2 label=(height=15pt "Top 10 Sources") minor=none;
ods region x=5.5 in y=0.3 in height=3.5 in width=5 in;
proc gchart data=topsources;
hbar source/ sumvar= _TEMA001 maxis=axis1 raxis=axis2;
run;quit;
axis1 label=(angle=90 "Tweet Owner")  minor=none;
axis2 label=(height=15pt "Top 10 Influencers") minor=none;
```

```
ods region x=0.25 in y=4.75 in height=3.5 in width=5 in;
proc gchart data=topowners;
hbar tweet_owner/ sumvar= _TEMA001 maxis=axis1 raxis=axis2;
run;quit;
axis1 label=(height=15pt "Tweets  per day")  minor=none ;
axis2 label=( angle=90 "Tweets") minor=none major=(n=4);
ods region x=5.5 in y=4.75 in height=3.5 in width=5 in;
proc gchart data=&dataset;
pattern1 color=red value=r3;
vbar pubdate /discrete outside=freq maxis=axis1 raxis=axis2 ;
run;quit;
ods layout end;
ods pdf STARTPAGE=NOW;
ods layout start;
ods region x=0 in y=0 in height=8.5 in width=11 in;
proc gslide;
title1 h=17pt j=left ' Tweet Report-Top Retweets' lspace=.25in;
run;quit;
ods region x=0 in y=0.3 in height=7.5 in width=10 in; title1;
proc sql outobs=20;
select Text 'Tweet',count(*) 'Retweets'
from &dataset._rt
group by Text
order by 2 desc;
quit;
ods layout end;
ods pdf close;
goptions reset=all;
ods listing;
%EndMacro:
```

```
%put Collected Tweets:&twtcount, Collected Retweets:&retwtcount;

%mend gettweet;

options mprint nomlogic nosymbolgen;

%let path=%nrstr(C:\Documents and Settings\Administrator\Desktop\New folder);

%put &authorization;

%let authorization=%nrstr("Authorization: cHJhdmthcm5hdmF0QGdtYWlsLmNvbQ0KdHVlc2RheTkj==");

%GetTweet(WORDS=Steve jobs, CODE= &authorization, PATH=&path);
```